RESEARCH ARTICLE OPEN ACCESS

Convergence of Big Data and Cloud Computing

Karambir kaur, Harinderjit Kaur, Surbhi
Department of Computer Science and Engineering,
Punjab Institute of Technology
Kapurthala – India

ABSTRACT

Big Data is data that either is too large, grows at an enormous rate, or does not fit into traditional architectures. Within such, precious information from data can be discovered through data analysis. With the emergence of cloud computing services, big data processing and analysis has become a less costly assignment. Clouds are being used to deal with the big data to effectively store and utilize the unstructured or semi-structured data of the organizations. In this paper, the current trends of Big Data, its analysis and how the data is transferred and stored in clouds, is explored. From the perspective of how big data is related to cloud computing, we present the key issues of big data storage and management, and then to solve the issues, cloud computing platform, cloud database, hadoop modules, and data storage scheme are explained.

Keywords:- Cloud Computing; Big Data; Hadoop

I. INTRODUCTION

1.1 Big Data

History of Big Data

- The term "Big Data" was introduced by SGI's chief scientist, John r. Masey, as early as in 1998 [1].
- In 2001 research report, META group (Gartner) analyst Doug Laney defined big data as three dimensional: volume, velocity and variety.
- Later in 2012, Gartner updated definition of big data as high volume, high velocity, and high variety [9].
- The large Big Data revolution is still ahead of us so there will be a lot of change in the coming years. Let the Big Data era begin!

Big data is a term used to describe a massive collection of both structured and unstructured data, availability and processing of large volumes of streaming data in real-time [1]. Big data is being generated by everything like digital processing and social media exchange around us at all times. Sensors, Systems and mobile devices transmit it. Big data is coming from multiple sources at an alarming velocity, volume and variety [4]. Every day we are creating more than 2 quintillion of data- so much that 90% of the data in the world has been created in the last two years alone. This data comes from every field like-sensors used to gather climate information, images, videos uploaded to social media sites, data gathered by government and administration, transaction records by bankers, etc.

Initially, from the recorded time until 2003, we created 5 billion GB of data. In 2011, the same amount of data was created in every two days. In 2013, the same amount of data is produced every 10 minute. A new field of science – Data Science, which is combination of mathematics, CS and art, is emerging. There is a huge demand of data scientist worldwide who apply methods of data science to better analyse these gigantic databases. As far as commercial application of big data analytics is concerned, market revenue was $28 billion in 2014 and expected to boost up to $50 billion in year 2017.

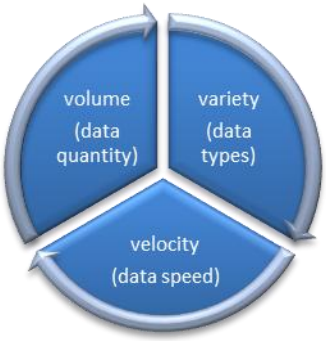

Figure 1. Characteristics of Big Data

1.2 Big Data Analytics

The real value of big data is when it is analyzed. Big data analytics refers to process of collecting, organizing and analysing large sets of data to discover patterns,

derive meanings, significant factor for decision, and ultimately the capability to respond to the world with greater agility. Big data analytics is a set of advanced technologies designed to work with large volumes of complex and heterogeneous data. It uses refined quantitative methods such as machine learning, neural networks, robotics, computational mathematics, and artificial intelligence to explore the data and to discover interrelationship and patterns [2].

1.3 Cloud Computing

History of Cloud Computing

- The term Cloud Computing was firstly coined in 199
- The term was actually introduced in 2006, when large companies like Google and Amazon began using "Cloud Computing" to describe the new trends in which people were increasingly accessing software, applications, computer power, and files over the Web instead of on their Personal computers [10].

Cloud computing, in computer networking, is computing that involves a large number of computers connected through a network such as the Internet [1]. Cloud computing, often referred to as simply "the cloud," is delivery of on-demand computing resources—everything from applications to data centres—over the Internet on a pay-per-use basis. The key target of cloud computing is to share resources, which consist of infrastructure, platform, software, and business process. From a cloud offering insight, Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), and Business Process as a Service (BPaaS) are typical service delivery types in cloud computing [2].

TABLE 1 EXAMPLES OF CLOUD COMPUTING SERVICES

Cloud Computing Services	**Examples**
Social Networking	Facebook, Twitter, Myspace, LinkedIn.
Email	Microsoft's Hotmail and Windows Live Mail.
Document/Spreadsheet/Other hosting Services	Google Docs allow users to share and edit their documents online from anywhere.
Backup Services	JungleDisk, Carbonite and Mozy allow public to automatically back up all their data to servers spread around the country or world [8].

II. RELATION BETWEEN BIG DATA AND CLOUD COMPUTING

Cloud Computing and Big Data are conjoined. The Big Data processing for organizations of all sizes are empowered by the cloud by relieving a number of problems. Cloud Computing democratizes and empowers big data – any enterprise can now work with unstructured data at a large scale. Cloud computing by bringing the big data analyses to the masses has provided businesses with affordable and flexible to immense amounts of computing resources on demand.

2.1 Cloud and Big Data: A Compelling Combination

Cloud delivery models offer a great flexibility, enabling organisations to evaluate the best approach to each business user's requirements. For example, organizations that already support a private cloud environment can add big data analytics to their in-house offerings, use a cloud services provider, or build a hybrid cloud that protects certain sensitive data in a private cloud, but takes advantage of important external data sources and applications provided in public clouds. Using cloud infrastructure to analyse big data, makes sense because:

- **It coerces a need for efficient, cost-effective infrastructure.**

The resources to support distributed computing models in-house typically reside in large and midsize data centres. Private clouds can offer a more efficient, cost-effective model to implement analysis of big data domestically, while expanding internal resources with public cloud services. The hybrid cloud option enables companies to use on-demand storage space and computing power via public cloud services for certain analytics initiatives (for example, short-term projects), and provide added capacity and scope as required [2].

- **Big Data can blend Internal and External Sources.**

While enterprises often keep their most sensitive and private data in-house, huge volumes of big data (owned

by the organization or created by third-party and public providers) may be located externally—some of it already in a cloud environment. Moving relevant data sources behind your firewall can be a significant assurance of resources. Analyzing the data where it resides—either in internal or public cloud data centres and client devices—often makes more sense [2].

2.2 The Cloud as an Enabler for Big Data Analytics

Two big IT initiatives are currently on top of mind for organizations throughout the world: big data analytics and cloud computing. An important aim of cloud computing is to deliver computing as a solution for managing big data, such as large scale multi-media and high dimensional data sets. As a delivery model for IT services, it has the potential and capability to enhance business awareness and productivity by enabling greater efficiencies and reducing costs.

Both technologies continue to grow. Organizations are moving beyond questions of what and how to store big data and finally how to derive meaningful analytics that respond to meet real business needs. As cloud computing continues to evolve, a great number of enterprises are building efficient and agile cloud environments, and cloud providers continue to expand service offerings. After this, it makes sense, that IT organizations should look to cloud computing as the structure to support their big data projects [2].

2.3 Big Data Trends

What makes cloud computing such a cost-effective delivery model for big data analytics? How are big data and cloud technologies converging to make big data analytics in clouds a sound option? For big data analytics:

- **Data is becoming more valuable**.

Today the conversation is shifting from "What data should we store?" to "What can we do with the data?" Organisations are looking to unlock data's hidden potential and deliver competitive advantage. Companies must find new approaches to processing, managing, and analyzing their data. The scope of big data analytics continues to expand.

- **Data Analytics is moving from batch to real time.**

Real time supports predictive analytics. Predictive analytics enables organizations to move to a future-oriented view of what's ahead and offers organizations many exciting opportunities for driving value from big data. Real-time data provides the prospect for fast, accurate, and flexible predictive analytics that quickly adapt to changing business conditions. The faster you analyze your data, the more sensible the results, and the greater it's predictive value [2].

2.4 Challenges of Big Data

- **Collection**

One of the major problems with big data is its huge size. The world's data is growing at an alarming velocity. The major problem with this system would be getting the data into the cloud to begin processing. Using standard Internet connections to upload the data to the cloud would be a bottleneck in the process. New techniques need to be investigated and developed to increase the efficiency of data movement into the cloud as well as across clouds [3].

- **Storage**

A significant problem with handling big data is the type of storage. Using a cloud approach, the traditional database is not currently suited to take advantage of the cloud's scalability. Current systems that exist handle scalability but do so at the expense of many of the advantages the relational model provides. New systems need to carefully take into account the requirement for these features while also providing a scalable model.

- **Analysis**

The major reason behind the need for handling big data is to be able to obtain value from data analysis. Analytic techniques and methods need to be further researched to develop techniques that can be able to process rising data sets. Simplification of the analysis process of big data is a major goal behind big data [3].

- **Security**

There is a great focus on two main problems when securing the large data systems. The first is to secure these systems such that there is a limited amount of overhead introduced so that performance will be greatly unaffected. More research and development needs to be conducted on securing data throughout the data analysis process. Unlike most RDBMS, No SQL security is largely relied on outside of the database system.

Research on the types of attacks that are possible on these new systems would be favourable [3].

III. BIG DATA IN THE CLOUD

3.1 Data Transfer

To take advantage of the cloud for big data analysis, data must first be loaded on the cloud. To address the issue of uploading big data to the cloud, a number of approaches to WAN optimization have been recognized. The techniques include: Compression, data deduplication, caching, and protocol optimization.

Compression can provide some benefits for WAN optimizations. These benefits would be dependent on the type of data that is being compressed. For example, plain text data would be more compressible than encrypted data [3]. Another technique that can be used to reduce the size of data being transported is data deduplication. Data deduplication looks at data both at the file and block level. When there are duplicates that exist, they are replaced with a pointer to the other copy. This technique is also called redundancy elimination. Other techniques are targeted towards protocol optimizations. In this protocol, one TCP and UDP port are used for session control and data transfer [3].

Some major cloud vendors now offer a service in which clients can ship physical media to the data centre, where it can be uploaded, and eliminating overly long data transfer times. Bulk imports are especially useful when data is first ported to the cloud or for backup and offsite storage. The fees for this service differ, and some cloud providers will also download data from the cloud and ship it via physical media.

TABLE II: COMPANIES TRANSFERRING DATA TO CLOUD

Company Name	Work done
AWS Import/Export	Accelerates transferring large amounts of data between the AWS cloud and portable storage devices that clients send to Amazon.
Google Cloud Storage Offline Disk Import	The service gives clients the option to load data into Google Cloud Storage directly by sending Google physical hard drives that it loads into an empty Cloud Storage bucket. This approach might be faster or less costly than transferring data over the Internet.
Rackspace's Bulk Import to Cloud Files	Is a service that lets clients send Rackspace physical media to be uploaded directly at the data centres, where migration specialists connect the device to a workstation that has a direct link to Rackspace's Cloud Files infrastructure [5].

3.2 Data Storage & Management

Big data has changed the architecture of systems for data storage. The two major factors for this change are in the need to be highly scalable and flexible enough to effectively handle big data. For storage, distributed systems like the Google File System were designed to use commodity clusters for storage that is both reliable and efficient. In this system, data is stored as file blocks of 64MB across the nodes of the cluster. On top of GFS, MapReduce is used for processing data across the nodes. It is more efficient to push computations to where the data resides rather than the opposite. MapReduce takes advantage of the distributed architecture of the file system by sending jobs to the nodes on the cluster where the data resides. There has been a considerable amount of research on the MapReduce concept developed by Google for processing large data sets. This can be largely attributed to two reasons. The simplicity of the functions for processing and the challenges it handles (replication, storage, etc.) [3].

Big Data Architecture: In this section, we mainly discuss big data architecture from three key aspects:

- **Distributed File System**

TABLE 3 EXAMPLES OF DFS

Examples	Description
Google File System (GFS)	is a chunk-based distributed file system that supports fault-tolerance by data partitioning and replication. As a storage layer of Google's cloud computing platform, it is used to read input and store output of MapReduce.
Hadoop	Its data storage layer called Hadoop Distributed File System (HDFS), which is an open-source equivalent of GFS.
Amazon Simple Storage Service (S3)	is an online public storage web service offered by Amazon Web Services. This file system is targeted at clusters hosted on the Amazon Elastic Compute Cloud server-on-demand infrastructure.

- **Non-structured and Semi-structured Data Storage**

With the success of the Web 2.0, more and more IT companies have increasing needs to store and analyze

the ever growing data, such as search logs, crawled web content, and click streams, usually in the range of petabytes, collected from a variety of web services. However, web data sets are usually non-relational or less structured and processing such semi-structured data sets raises another challenge. Moreover, simple distributed file systems mentioned above cannot satisfy service providers like Google, Yahoo, Microsoft and Amazon. All providers have their purpose to serve probable users and own their relevant state of- the-art of big data management systems in the cloud environments [6].

TABLE 4 PLATFORMS FOR MANAGING NON-STRUCTURED AND SEMI-STRUCTURED STORAGE

Example	Description
Bigtable	is a distributed storage system of Google for managing structured data that is designed to scale to a very large size (petabytes of data) across thousands of commodity servers. It provides clients with a simple data model that supports active control over data layout and format.
PNUTS	is an enormous scale hosted database system designed to support Yahoo!'s web applications. Upon PNUTS, new applications can be built very easily and the overhead of creating and maintaining these applications is nothing much.
Dynamo	is a highly available and scalable distributed key/value based data store built for supporting internal Amazon's applications. It provides a simple primary-key only interface to meet the needs of these applications [6].

- **Open Source Cloud Platform**

The main idea behind data centre is to influence the virtualization technology to maximize the utilization of computing resources. Therefore, it provides the basic ingredients such as storage, CPUs, and network bandwidth as a commodity by expert service providers at low cost. Amazon Web Services (AWS), Open nebula, Cloud stack and Open stack are the most popular cloud management platforms for infrastructure as a service (IaaS) [6].

TABLE 5 CLOUD MANAGEMENT PLATFORMS

Cloud Management Platforms	Description
AWS9	has enormous usage in elastic platform. It is very easy to use and only pay-as-you-go.
Eucalyptus	works in IaaS as an open source. It uses virtual machine in controlling and managing resources.
OpenNebula	has integration with various environments and offer the richest features, flexible ways to build private, public or hybrid clouds.
CloudStack	users can take full advantage of cloud computing to deliver higher efficiency and faster deployment of new services to the users. It is one of the Apache open source projects [6].

The most popular implementation, Hadoop consists of two components, the MapReduce engine and Hadoop Distributed File System. However the fact that the MapReduce paradigm is essentially both index and schemaless has been a point of conflict against the framework. This has lead to the development of several systems built on top of the Hadoop core components to address these problems.

3.3 Hadoop

Hadoop is a framework of tools that supports reliable distributed computing and processing of large datasets across large clusters of computers. Hadoop is an open source implementation for Google MapReduce, written in java and maintained under Apache License. Based on Google's white papers for GFS and MapReduce, there are three major components: HDFS, the MapReduce engine, and the utilities for the other Hadoop Modules. Hadoop makes it easier to store, process and analyze lot of data on commodity hardware. There are several Hadoop modules:

Figure 2. Hadoop Ecosystem [7]

TABLE 6 HADOOP MODULES

Name	Description
Hive	Apache Hive is a data warehousing

	system used with Hadoop for querying, summarization, and analysis of data stored in Hadoop. Queries are expressed in the "SQL-like" Hive Querying Language. Hive organizes data into tables, partitions and buckets. A table logically consists of rows and columns.
Pig	Apache Pig is described as a platform for the analysis of large datasets. Like Hive, it uses a high level language that is compiled into MapReduce programs that are executed on Hadoop. Pig's high-level language is called Pig Latin.
Hbase	Hbase is a distributed, column-oriented NoSQL database that operates on top of the HDFS and is fault tolerant. Modelled Google's BigTable, Hbase provides a distributed data store that is highly scalable with consistent reads and writes. Data is stored as indexed Store files on HDFS.
ZooKeeper	Provide coordination services such that synchronization can be enabled throughout a Hadoop cluster, achieves this by maintaining objects containing information and namespaces in-memory. This information is kept across distributed ZooKeeper servers that would retrieve the client applications that require them [3].

3.4 Cassandra

It was initially developed by facebook and built on Amazon's Dynamo and Google's BigTable, is a distributed column-oriented database system for storing and managing very large amounts of structured data widen across many commodity servers, while providing highly available service with no single point of failure [12]. Cassandra is a distributed column-oriented database. The smallest piece of data is the column consisting of a name, value and timestamp. A row contains multiple columns, column families contain rows, and key spaces contain column families. Column families are stored in separate files. Data is separated into partitions across the nodes in the distributed database [3]. The Apache Cassandra database is an accurate choice when you need scalability and high availability without compromising performance. Linear scalability and demonstrated fault-tolerance on commodity hardware or cloud infrastructure make it the perfect platform for mission-critical data [11].

Many large companies have successfully deployed and benefited from Apache Cassandra including Adobe, Comcast, eBay, Rackspace, Netflix, Twitter and Cisco. The larger production environments have hundreds of TB of data in clusters of more than 300 servers. Cassandra is available under the Apache license [12].

3.5 Voldemort

Developed by LinkedIn, Voldemort is a highly scalable, distributed key-value data store. Data is automatically replicated and partitioned between the nodes in the distributed system, so each server contains only a subset of the total data. Server failure is handled in transparent manner. Each node is independent of the others such that there is no single point of failure. Read and write access is restricted to key-value access. As such, there are only three types of queries available: get, put and delete. This simplicity provides predictability in the performance of queries [3].

IV. CONCLUSION

This paper described a systematic flow of survey on big data transfer, storage and management in the perspective of cloud computing. Here, challenges for big data, then cloud storage and computing architecture, popular parallel processing framework are respectively discussed. Big Data is becoming a new way for exploring and discovering interesting, valuable information. The volume of data that exists is constantly enlarging such that the majority of data that exists has been created in just the past few years. With that in mind, the "big data" of tomorrow can be expected to be magnitudes larger than it is by today's standards. As a result, the problems of big data will only become more prominent and dominant in the future as solutions are being developed to meet the emerging needs.

Hadoop has become a common solution for processing large amounts of data. However, Hadoop uses a batch-processing approach and does not provide adequate solutions for real-time ad hoc querying needs. Solutions such as Pig and Hive provide a means of simplifying

querying, yet at bottom they are still using MapReduce jobs to query Hadoop. Future development is expected to focus on systems that provide real-time, ad hoc querying capabilities over large scale data.

REFERENCES

[1] LiMa, Mingfeng Jiang, "Chances and Challenges Confronting Securities Industry and the Counter measures in Big Data and Cloud Computing Era", The 9th International Conference on Computer Science & Education (ICCSE 2014) august 22-24, 2014

[2] Big data in the cloud: converging technologies http://www.intel.in/content/dam/www/public/us/en/documents/product-briefs/big-data-cloud-technologies-brief.pdf

[3] Sanjay P. Ahuja & Bryan Moore, "State of Big Data Analysis in the Cloud", Published by Canadian Center of Science and Education, May 22, 2013.

[4] http://www.ibm.com/big-data/us/en/

[5] http://gcn.com/articles/2013/08/05/bulk-import.aspx

[6] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li, "Big Data Processing in Cloud Computing Environments", International Symposium on Pervasive Systems, Algorithms and Networks, 2012.

[7] https://technocents.wordpress.com/2014/03/24/hadoop- ecosystem/

[8] Hirdesh Shivhare, Nishchol Mishra, Jitendra Agarwal, Sanjeev Sharma, "Cloud Computing and Big Data", International Conference on Cloud, Big Data and Trust 2013, nov 15, 2013.

[9] http://www.hcltech.com/sites/default/files/big-data-timeline.jpg

[10] http://www.technologyreview.com/news/42597 0/who-coined-cloud-computing/

[11] http://cassandra.apache.org/

[12] http://planetcassandra.org/what-is-apache-cassandra/